PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-147205
(43)Date of publication of application : 07.06.1996

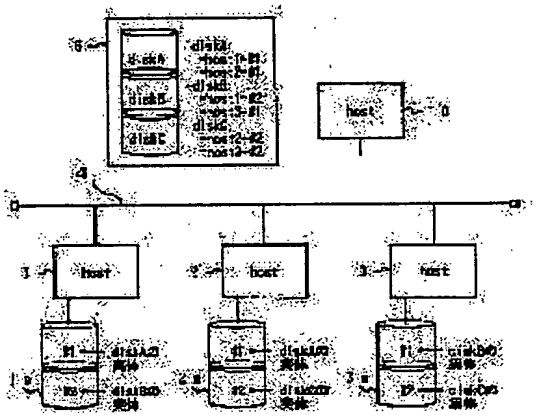
(51)Int.Cl. G06F 12/00
G06F 12/00

(21)Application number : 06-285150
(22)Date of filing : 18.11.1994

(71)Applicant : NEC CORP
(72)Inventor : FUKUSHIMA SETSU

(54) DISK SHARING SYSTEM

(57)Abstract:
PURPOSE: To provide the disk sharing system with which the capacity of physical disks can be enlarged and access control to a shared disk can be performed with high reliability.
CONSTITUTION: The shared disk is constituted by providing physical disks A, B and C to be the components of a virtual disk from host machines 1-3 to a managing host machine 10. An up access request issued by the respective host machines 1-3 by providing semaphores to the physical disks A, B and C is controlled by the semaphore function of the managing host machine 10 by executing a process inside the managing host machine 10. The process gets rest until the semaphore is acquired and at the time point when the semaphore is acquired, a down access request is issued to the host machine equipped with the relevant physical disk. When access is completed, the semaphore is recovered and the processing, which gets rest in a recovery waiting state, is woken up. With this procedure, exclusive control is performed.



LEGAL STATUS

[Date of request for examination] 18.11.1994
[Date of sending the examiner's decision of rejection] 17.02.1999
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]
[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平8－147205

(43)公開日　平成８年(1996)６月７日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　12/00	５３５　Ａ	7623－5B		
	５４５　Ａ	7623－5B		

審査請求　有　　請求項の数３　ＯＬ　（全 10 頁）

(21)出願番号　特願平6－285150

(22)出願日　平成６年(1994)11月18日

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号

(72)発明者　福島　節
東京都港区芝五丁目７番１号　日本電気株式会社内

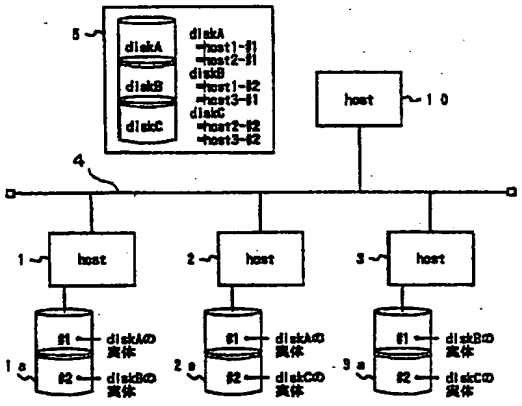
(74)代理人　弁理士　後藤　洋介　（外２名）

(54)【発明の名称】　ディスク共有システム

(57)【要約】

【目的】　物理ディスクの大容量化を図り得ると共に、共有ディスクへのアクセス制御を信頼性高く行い得るディスク共有システムを提供すること。

【構成】　ホストマシン１，２，３から仮想ディスク５の構成要素となる物理ディスクＡ，Ｂ，Ｃを管理ホストマシン１０に対して提供して共有ディスクを構成する。各ホストマシン１，２，３が物理ディスクＡ，Ｂ，Ｃに対してセマフォを提供して発行した上がりアクセス要求は、管理ホストマシン１０内のプロセスを実行することで管理ホストマシン１０のセマフォ機能により制御される。プロセスはセマフォを獲得するまで休眠し、セマフォを獲得した時点で該当する物理ディスクを持つホストマシンに対して下がりアクセス要求を発行する。アクセスが完了したならセマフォを回復し、回復待ち状態で休眠していたプロセスを目覚めさせる。このような手順をとることで排他制御を行う。



1

【特許請求の範囲】

【請求項１】　ネットワーク上に分散されて配置されると共に、それぞれの物理ディスクで冗長化された仮想ディスクを共有ディスクとして構成する複数のホストマシンと、前記ネットワーク上に配置されると共に、前記共有ディスクへのアクセス要求に従ってセマフォ機能のプロセスを実行することで前記複数のホストマシンを排他制御して管理する管理ホストマシンとを含むことを特徴とするディスク共有システム。

【請求項２】　請求項１記載のディスク共有システムにおいて、前記複数のホストマシンは、それぞれの物理ディスクの領域に別のホストマシンと同じ内容の領域を持つと共に、前記仮想ディスクに対してセマフォを提供して上がりアクセス要求を発行するものであり、前記管理ホストマシンは、前記仮想ディスク及び前記それぞれの物理ディスクを関連付けるための関連情報を有するテーブルを持つと共に、前記セマフォの獲得により前記上がりアクセス要求を受信したときに前記プロセスを生成することを特徴とするディスク共有システム。

【請求項３】　請求項２記載のディスク共有システムにおいて、前記管理ホストマシンは、前記セマフォを獲得するまでの間，前記複数のホストマシンのうちの該当するものに対して前記プロセスの生成を休眠させて回復待ち状態とし、該セマフォの獲得後，該複数のホストマシンのうちの該当するものに対して下がりアクセス要求を発行し、更に該下がりアクセス要求の発行完了後に該セマフォを回復して該回復待ち状態の該プロセスの生成を目覚めさせることを特徴とするディスク共有システム。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、分散された物理ディスクの冗長化と排他制御とを行うディスク共有システムに関する。

【０００２】

【従来の技術】従来、この種のディスク共有システムで仮想ディスクの排他制御を行う場合、例えば特開昭５９－０５８５７１号公報に開示された複数の論理ボリュームを持つ磁気ディスク装置の制御方式に記載されているように、仮想ディスクの構成要素となる物理ディスクを１台のホストマシン内に存在させ、この１つの物理ディスクを複数の仮想ディスクに分割してそれぞれ仮想ディスクをロック状態する方法や、或いは特開昭６３－２０８９２４号公報に開示された多重化ボリューム制御方式に記載されているように、多重化された物理ディスクを１つの入出力装置内に存在させ、仮想ディスクに対するロック要求を順番に登録して実行する方法等が施行されている。

【０００３】

【発明が解決しようとする課題】上述したディスク共有システムの場合、仮想ディスクの構成要素である物理ディスクが１台のホストマシンによって提供されるため、そのホストマシンの持つ容量が提供可能な容量の上限となってしまう。

【０００４】又、仮想ディスクを提供するホストマシンのシステムが支障を来して使用不能になった場合、そのホストマシン内の物理ディスクには冗長部分も含めてアクセス不可能となるため、その物理ディスクを共有ディスクとする他のホストマシンを含めてディスク共有システム全体の停止を余儀無くさせられてしまう。

【０００５】そこで、ディスク共有システムで仮想ディスクを提供するための物理ディスクを有するホストマシンを複数にして仮想ディスクを構成することも可能であるが、この場合には各ホストマシン間で相互に物理ディスクの利用状況を交換しなければならず、非常に複雑な手順が必要となるという難点がある。

【０００６】本発明は、このような問題点を解決すべくなされたもので、その技術的課題は、物理ディスクの大容量化を図り得ると共に、共有ディスクへのアクセス制御を信頼性高く行い得るディスク共有システムを提供することにある。

【０００７】

【課題を解決するための手段】本発明によれば、ネットワーク上に分散されて配置されると共に、それぞれの物理ディスクで冗長化された仮想ディスクを共有ディスクとして構成する複数のホストマシンと、ネットワーク上に配置されると共に、共有ディスクへのアクセス要求に従ってセマフォ機能のプロセスを実行することで複数のホストマシンを排他制御して管理する管理ホストマシンとを含むディスク共有システムが得られる。

【０００８】又、本発明によれば、上記ディスク共有システムにおいて、複数のホストマシンは、それぞれの物理ディスクの領域に別のホストマシンと同じ内容の領域を持つと共に、仮想ディスクに対してセマフォを提供して上がりアクセス要求を発行するものであり、管理ホストマシンは、仮想ディスク及びそれぞれの物理ディスクを関連付けるための関連情報を有するテーブルを持つと共に、セマフォの獲得により上がりアクセス要求を受信したときにプロセスを生成するものであるディスク共有システムが得られる。

【０００９】更に、本発明によれば、上記ディスク共有システムにおいて、管理ホストマシンは、セマフォを獲得するまでの間，複数のホストマシンのうちの該当するものに対してプロセスの生成を休眠させて回復待ち状態とし、該セマフォの獲得後，該複数のホストマシンのうちの該当するものに対して下がりアクセス要求を発行し、更に該下がりアクセス要求の発行完了後に該セマフォを回復して該回復待ち状態の該プロセスの生成を目覚めさせるディスク共有システムが得られる。

【００１０】

【実施例】以下に実施例を挙げ、本発明のディスク共有

システムについて、図面を参照して詳細に説明する。図1は、本発明の一実施例に係るディスク共有システムの基本構成を示したブロック図である。

【0011】このディスク共有システムは、ネットワーク4上に分散されて配置された3台のホストマシン1，2，3と、ネットワーク4上に配置された管理ホストマシン10とから成っている。各ホストマシン1，2，3は、それぞれの物理ディスク1a，2a，3aで冗長化された仮想ディスク5を共有ディスクとして構成しており、管理ホストマシン10は共有ディスクへのアクセス要求に従ってセマフォ機能のプロセスを実行することで各ホストマシン1，2，3を排他制御して管理する。仮想ディスク5はディスクを二重化する構成の物理ディスクA，B，Cから構成され、各物理ディスク1a，2a，3aはそれぞれ物理ディスクAの実体及び物理ディスクBの実体，物理ディスクAの実体及び物理ディスクCの実体，物理ディスクBの実体及び物理ディスクCの実体により構成されている。

【0012】ここで、仮想ディスク5上の各物理ディスクA，B，Cの実体に関し、例えば物理ディスクAの構成はホスト1－#1，ホスト2－#1に二重化され、物理ディスクBの構成はホスト1－#2，ホスト3－#1に二重化され、物理ディスクCの構成はホスト2－#2，ホスト3－#2に二重化されるという具合に必ず異なるホストマシン上で多重化する。即ち、ここでは各ホストマシン1，2，3がそれぞれの物理ディスク1a，2a，3aの領域に別のホストマシンと同じ内容の領域を持つ構成をとることにより、仮想ディスク5を共有するホストマシン1，2，3はそれぞれがローカルに持つディスク容量以上の大きさのパーティションを冗長化構成で持つことができる。これにより、各ホストマシン1，2，3の何れかのシステムがダウンしてしまっても、稼働中のマシンの物理ディスクより仮想ディスク5を構成することができるので、ディスク共有システム全体が停止されること無く継続して作業を行うことが可能となる。

【0013】図2は、このディスク共有システムによる動作として冗長化された仮想ディスク5へのアクセス要求手順を示したものである。ここでは、各ホストマシン1，2，3が仮想ディスク5に対してセマフォを提供して上がりアクセス要求を発行し、管理ホストマシン10が仮想ディスク5及びそれぞれの物理ディスク1a，2a，3aを関連付けるための関連情報を有するテーブルを持つと共に、セマフォの獲得により上がりアクセス要求を受信したときにプロセスを生成する。管理ホストマシン10は、セマフォを獲得するまでの間，各ホストマシン1，2，3のうちの該当するものに対してプロセスの生成を休眠させて回復待ち状態とし、セマフォの獲得後，各ホストマシン1，2，3のうちの該当するものに対して下がりアクセス要求を発行し、更に下がりアクセス要求の発行完了後にセマフォを回復して回復待ち状態のプロセスの生成を目覚めさせる。このような手順で排他制御をファイル単位で行うと、プロセスの休眠／目覚めが管理され、物理ディスク（そのファイル）A，B，Cのセマフォが獲得される。

【0014】具体的に云えば、ホストマシン3よりアクセス要求を発行して物理ディスクAの実体1a′，2a′を読み出す場合、先ずホストマシン3が管理ホストマシン10に対してセマフォを提供して上がりアクセス要求W1を発行し、この上がりアクセス要求W1を受信した管理ホストマシン10がそれを実行するためのプロセスを生成する。ここでのプロセスは、物理ディスクAの関連情報を示すファイルAへのアクセス待ちプロセスであり、管理ホストマシン10が提供する各ファイルA，B，Cに関するセマフォ機能により制御される。これにより、プロセスはセマフォの獲得を試みるが、獲得できなかったら休眠する。

【0015】又、プロセスがファイルAのセマフォを獲得すると、ホストマシン3の物理ディスク3aに関する物理ディスクBの実体及び物理ディスクCの実体に対応する物理ディスク（ここでは物理ディスク1a，2aの双方が該当する）を持つホストマシン1，2に対してそれぞれ下がりアクセス要求W2を発行する。これにより、例えばホストマシン1，2からファイルAの実体1a′，2a′を読み出したリード結果W3がホストマシン3へ伝送される。この後、ホストマシン1，2からは管理ホストマシン10に対して上がりアクセス要求W4を発行して完了通知を行う。管理ホストマシン10では1つのプロセスが完了した時点でセマフォが回復し、そのセマフォ回復待ちのプロセスを目覚めさせる。

【0016】図3は、管理ホストマシン10内におけるセマフォ値の遷移を示したものである。ここでのセマフォ値は、それぞれ深さを3，ライトを含むモードでオープンされていない状態を2，アクセス可能状態を1，ロック状態を0としている。

【0017】図4～図9は、それぞれ管理ホストマシン10内で生成したプロセス1～6毎のセマフォ値の操作例を示したものである。各図においては、それぞれセマフォ値1及びセマフォ値2という二重化されたファイルに対応したセマフォ配列を持ち、共有ディスクである仮想ディスク5にアクセスするための処理手順を示している。

【0018】例えば図4は、ファイルのライトを含んだモードでのオープンを行う内容のプロセス1に関して、セマフォ配列がセマフォ値1＝2，セマフォ値2＝2の設定後に両方のセマフォの値を2減らす処理（失敗ならスリープ，オープンのためのロックを行う処理を含む）を行い、セマフォ配列がセマフォ値1＝0，セマフォ値2＝0の設定時にリード／ライトモードでオープンし、この後にセマフォの回復の処理（オープンのためのロッ

ク解除，アクセス可能な1にする処理を含む）を行ってからセマフォ配列をセマフォ値1＝1，セマフォ値2＝1に設定する例を示している。

【0019】図5は、ライトを含んだモードでオープンしたファイルのクローズを行う内容のプロセス2に関して、セマフォ配列がセマフォ値1＝1，セマフォ値2＝1の設定後に両方のセマフォの値を1減らす処理（失敗ならスリープ，クローズのためのロックを行う処理を含む）を行い、セマフォ配列がセマフォ値1＝0，セマフォ値2＝0の設定時にファイルをクローズし、この後にセマフォの回復の処理（クローズのためのロック解除，オープン可能な2に戻す処理を含む）を行ってからセマフォ配列をセマフォ値1＝2，セマフォ値2＝2に設定する例を示している。

【0020】図6は、ファイルへのライトを行う内容のプロセス3に関して、セマフォ配列がセマフォ値1＝1，セマフォ値2＝1の設定後に両方のセマフォの値を同時に1減らす処理（失敗ならスリープする処理を含む）を行い、セマフォ配列がセマフォ値1＝0，セマフォ値2＝0の設定時にディスク1，2にライトし、この後にディスク1のライト完了確認，セマフォ1の回復を行ってからセマフォ配列がセマフォ値1＝1，セマフォ値2＝0の設定を行い、更にこの後にディスク2のライト完了確認，セマフォ2の回復を行ってからセマフォ配列をセマフォ値1＝1，セマフォ値2＝1に設定する例を示している。

【0021】図7は、ファイルのリードを行う内容のプロセス4に関して、セマフォ配列がセマフォ値1＝1（2），セマフォ値2＝1（2）の設定後にセマフォ1の値を1減らすことが成功していれば所定の指定処理（失敗でもスリープしない，ディスクαのためのロックを行う処理を含む）を行い、セマフォ配列がセマフォ値1＝0（1），セマフォ値2＝1（2）の設定時にディスク1からリードし、この後にセマフォ1の回復の処理（ディスク1リードのためのロック解除，ブレイクを含む）を行ってからセマフォ配列がセマフォ値1＝1（2），セマフォ値2＝1（2）の設定にしてセマフォ2の値を1減らすことが成功していれば所定の指定処理（失敗でもスリープしない，ディスク2からリード，ブレイクを含む）を行うことを示している。

【0022】図8は、ファイルのリードオンリーモードでのオープンを行う内容のプロセス5に関して、セマフォ配列がセマフォ値1＝2（1），マフォ値2＝2（1）の設定後に両方のセマフォの値を1減らす処理（失敗ならスリープ，オープンのためのロック，リード／ライトでオープン済みであれば1とする処理を含む）を行い、セマフォ配列がセマフォ値1＝1（0），セマフォ値2＝1（0）の設定時にリードオンリーモードでオープンし、この後にセマフォの回復の処理（オープンのためのロック解除，オープンの前の値にする処理を含む）を行ってからセマフォ配列をセマフォ値1＝2（1），セマフォ値2＝2（1）に設定した例を示している。

【0023】図9は、リードオンリーでオープンしたファイルのクローズを行う内容のプロセス6に関して、セマフォ配列がセマフォ値1＝2（1），マフォ値2＝2（1）の設定後に両方のセマフォの値を1減らす処理（失敗ならスリープ，クローズのためのロックを行う処理を含む）を行い、セマフォ配列がセマフォ値1＝1（0），セマフォ値2＝1（0）の設定時にファイルをクローズし、この後にセマフォの回復の処理（クローズのためのロック解除，クローズ前の値にする処理を含む）を行ってからセマフォ配列をセマフォ値1＝2（1），セマフォ値2＝2（1）に設定した例を示している。

【0024】このような処理手順を管理ホストマシン10側で用いれば、冗長化されたファイル間の同期は維持され、排他制御も実現することができる。

【0025】

【発明の効果】以上に説明したように、本発明のディスク共有システムによれば、仮想ディスクを構成する物理ディスクの大容量化及び信頼性の向上を仮想ディスクの実体を分散させることによって実現しているので、ディスク資源の有効な活用が実現できると共に、障害が発生して特定のホストマシンにおけるシステムがダウンしたときにもディスク共有システム全体が停止せず、支障を来したホストマシン以外の他のホストマシンに関してはユーザが作業を継続できるようになる。又、冗長化されている共有ディスク（仮想ディスク）に対するアクセス要求に際しての排他制御を管理ホストマシンのセマフォ機能で実現しているので、共有ディスクの利用者は、冗長化された物理ディスク間の同期や二重書き込み，或いは物理ディスクの入出力（I／O）の衝突等を全く意識しなくて済むようになる。

【図面の簡単な説明】

【図1】本発明の一実施例に係るディスク共有システムの基本構成を示したブロック図である。

【図2】図1に示すディスク共有システムによる動作として冗長化された仮想ディスクへのアクセス要求手順を示したものである。

【図3】図1に示すディスク共有システムに備えられる管理ホストマシン内におけるセマフォ値の遷移を示したものである。

【図4】図1に示すディスク共有システムに備えられる管理ホストマシン内で生成したプロセス1のセマフォ値の操作例を示したものである。

【図5】図1に示すディスク共有システムに備えられる管理ホストマシン内で生成したプロセス2のセマフォ値の操作例を示したものである。

【図6】図1に示すディスク共有システムに備えられる

管理ホストマシン内で生成したプロセス３のセマフォ値の操作例を示したものである。

【図７】図１に示すディスク共有システムに備えられる管理ホストマシン内で生成したプロセス４のセマフォ値の操作例を示したものである。

【図８】図１に示すディスク共有システムに備えられる管理ホストマシン内で生成したプロセス５のセマフォ値の操作例を示したものである。

【図９】図１に示すディスク共有システムに備えられる

管理ホストマシン内で生成したプロセス６のセマフォ値の操作例を示したものである。

【符号の説明】

１，２，３　　ホストマシン

１ａ，２ａ，３ａ，Ａ，Ｂ，Ｃ　　物理ディスク

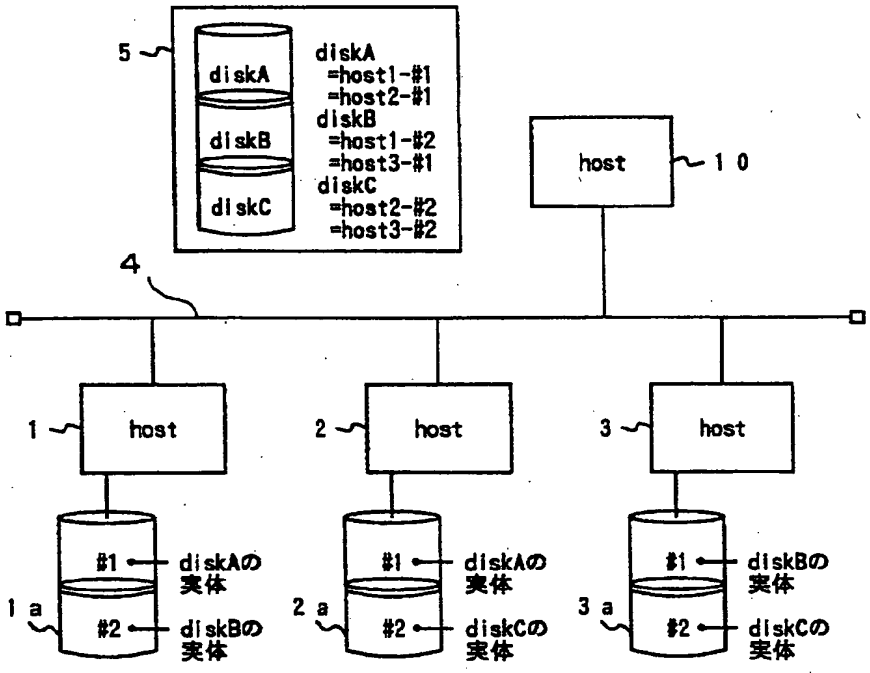
１ａ′，２ａ′　　物理ディスクＡの実体

４　　ネットワーク

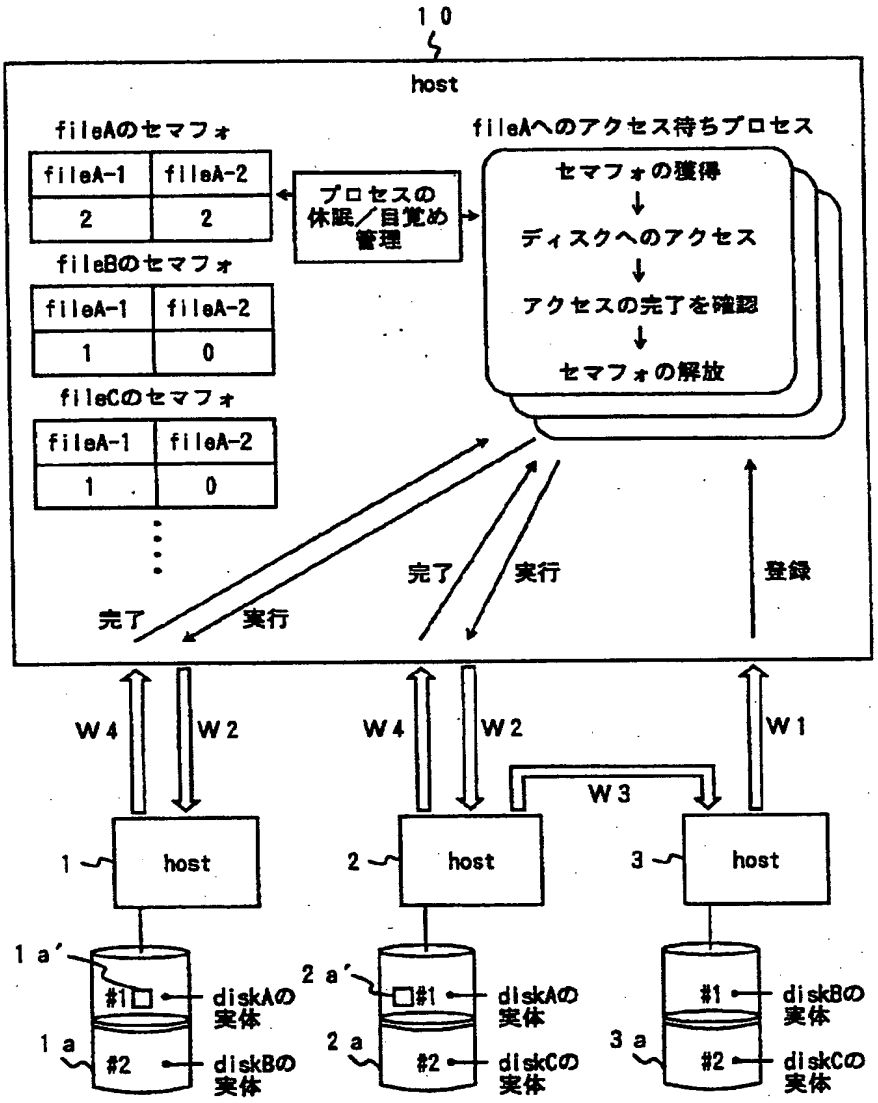
５　　仮想ディスク

１０　　管理ホストマシン

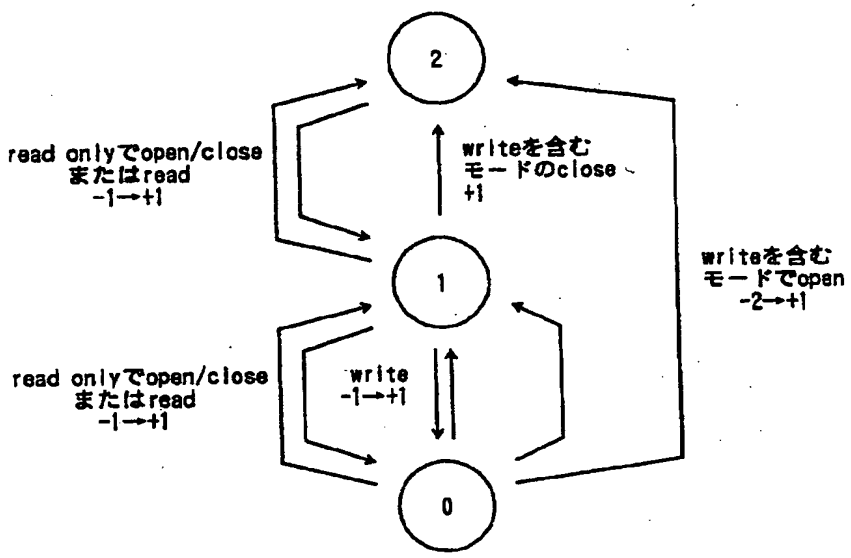
【図１】



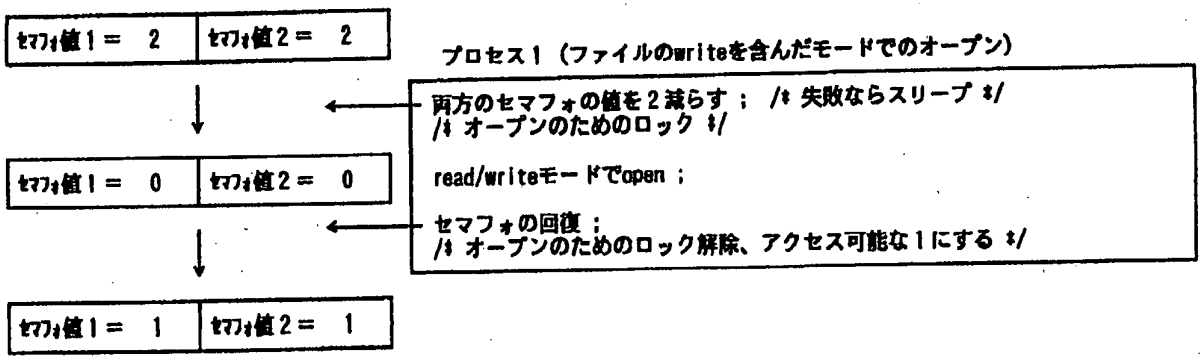
【図2】


10
host
fileAのセマフォ
fileA-1
fileA-2
2
2
fileBのセマフォ
fileA-1
fileA-2
1
0
fileCのセマフォ
fileA-1
fileA-2
1
0
プロセスの
休眠／目覚め
管理
fileAへのアクセス待ちプロセス
セマフォの獲得
ディスクへのアクセス
アクセスの完了を確認
セマフォの解放
完了
実行
完了
実行
登録
W4
W2
W4
W2
W3
W1
1
host
2
host
3
host
1 a'
#1
diskAの
実体
1 a
#2
diskBの
実体
2 a'
#1
diskAの
実体
2 a
#2
diskCの
実体
3 a
#1
diskBの
実体
#2
diskCの
実体

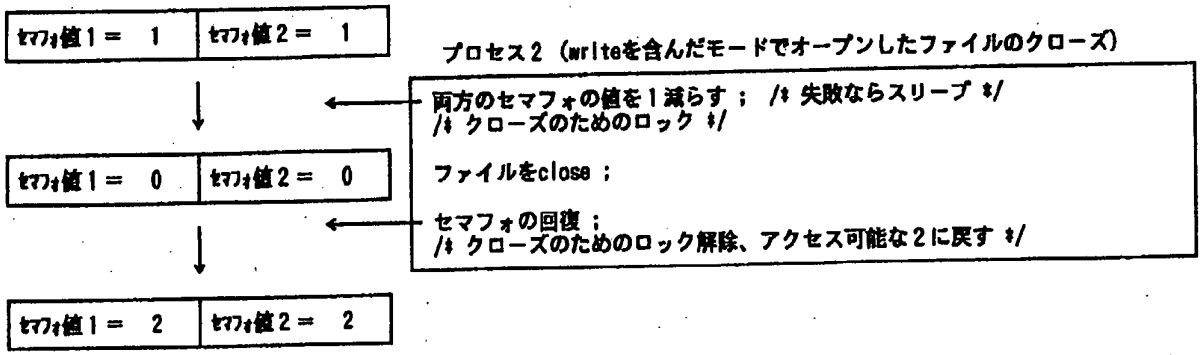
【図3】


2
1
0
read onlyでopen/close
またはread
-1→+1
writeを含む
モードのclose
+1
writeを含む
モードでopen
-2→+1
read onlyでopen/close
またはread
-1→+1
write
-1→+1

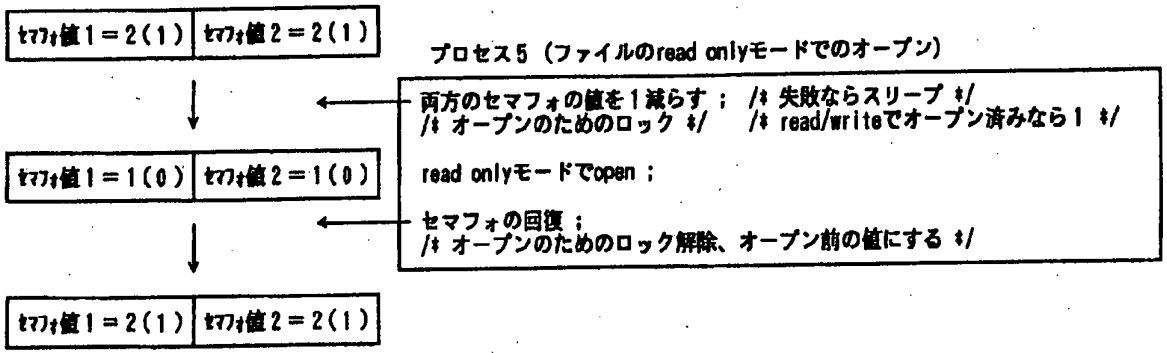
【図4】


セマフォ値1 = 2
セマフォ値2 = 2
プロセス1（ファイルのwriteを含んだモードでのオープン）
両方のセマフォの値を2減らす ； /* 失敗ならスリープ */
/* オープンのためのロック */
セマフォ値1 = 0
セマフォ値2 = 0
read/writeモードでopen ；
セマフォの回復 ；
/* オープンのためのロック解除、アクセス可能な1にする */
セマフォ値1 = 1
セマフォ値2 = 1

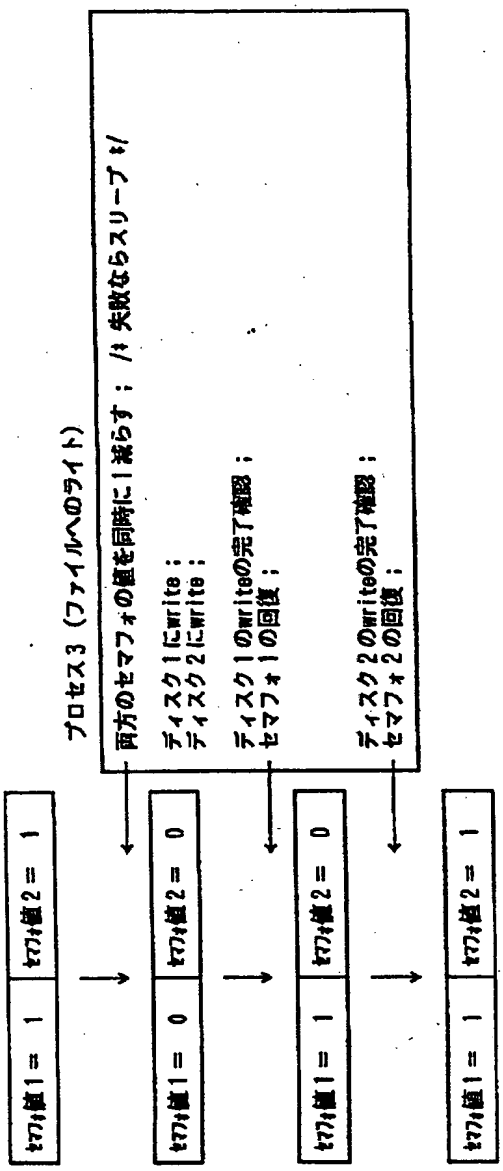
【図5】


セマフォ値1 = 1
セマフォ値2 = 1
プロセス2（writeを含んだモードでオープンしたファイルのクローズ）
両方のセマフォの値を1減らす ； /* 失敗ならスリープ */
/* クローズのためのロック */
セマフォ値1 = 0
セマフォ値2 = 0
ファイルをclose ；
セマフォの回復 ；
/* クローズのためのロック解除、アクセス可能な2に戻す */
セマフォ値1 = 2
セマフォ値2 = 2

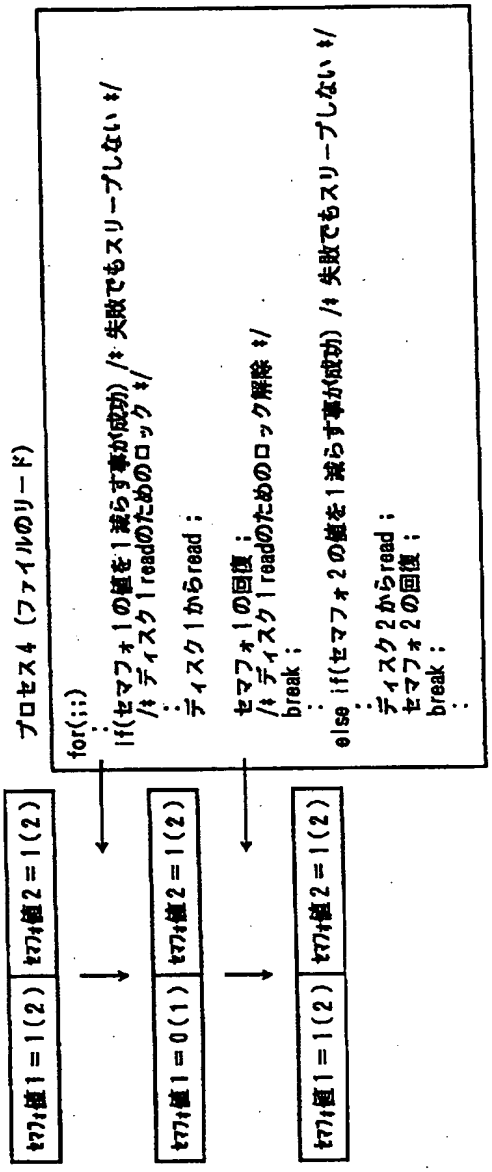
【図8】


セマフォ値1 = 2(1)
セマフォ値2 = 2(1)
プロセス5（ファイルのread onlyモードでのオープン）
両方のセマフォの値を1減らす ； /* 失敗ならスリープ */
/* オープンのためのロック */ /* read/writeでオープン済みなら1 */
セマフォ値1 = 1(0)
セマフォ値2 = 1(0)
read onlyモードでopen ；
セマフォの回復 ；
/* オープンのためのロック解除、オープン前の値にする */
セマフォ値1 = 2(1)
セマフォ値2 = 2(1)

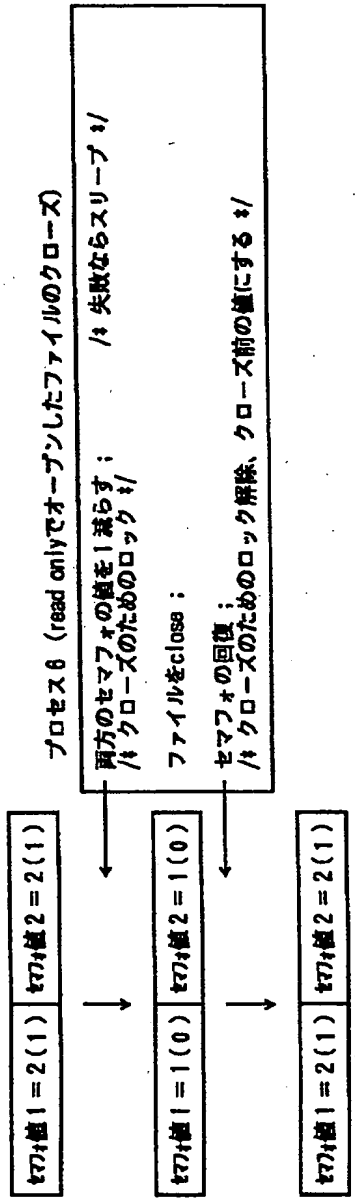
【図6】


プロセス3（ファイルへのライト）
両方のセマフォの値を同時に1減らす；　/* 失敗ならスリープ */
ディスク1にwrite；
ディスク2にwrite；
ディスク1のwriteの完了確認；
セマフォ1の回復；
ディスク2のwriteの完了確認；
セマフォ2の回復；
セマ値1＝1
セマ値2＝1
セマ値1＝0
セマ値2＝0
セマ値1＝1
セマ値2＝0
セマ値1＝1
セマ値2＝1

【図7】


プロセス4（ファイルのリード）
for(;;)
：
if(セマフォ1の値を1減らす事が成功)　/* 失敗でもスリープしない */
/* ディスク1readのためのロック */
：
ディスク1からread；
：
セマフォ1の回復；
/* ディスク1readのためのロック解除 */
break；
：
else if(セマフォ2の値を1減らす事が成功)　/* 失敗でもスリープしない */
：
ディスク2からread；
セマフォ2の回復；
break；
：
セマ値1＝1(2)
セマ値2＝1(2)
セマ値1＝0(1)
セマ値2＝1(2)
セマ値1＝1(2)
セマ値2＝1(2)

【図9】


プロセスB（read onlyでオープンしたファイルのクローズ）
両方のセマフォの値を1減らす；
/* クローズのためのロック */
/* 失敗ならスリープ */
ファイルをclose；
セマフォの回復；
/* クローズのためのロック解除、クローズ前の値にする */
セマ1値1＝2(1)
セマ1値2＝2(1)
セマ1値1＝1(0)
セマ1値2＝1(0)
セマ1値1＝2(1)
セマ1値2＝2(1)